REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
08-08-2012	Final Report	1-Nov-2011 - 31-Jul-2012

4. TITLE AND SUBTITLE
Final Report for Modeling Spatiotemporal Contextual Dynamics

5a. CONTRACT NUMBER
W911NF-11-1-0365

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

6. AUTHORS
Yun Fu

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
State University of New York (SUNY) at Buffalc
Sponsored Projects Services
The Research Foundation of SUNY on behalf of Univ at Buf
Buffalo, NY 14260 -

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
59118-CS-II.12

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
An important problem of visual understanding is how to recognize and predict human actions or imminent events from video. The ultimate intelligent systems should be able to detect/track suspicious subjects, predict actions and events, and raise alarms for emergencies before happening. From this STIR project, we have created a new algorithmic tool set of modeling spatiotemporal contextual dynamics. For low-level and middle-level visual representation, we proposed a class of Schatten norm based discriminative metrics, locality-constrained low-rank

15. SUBJECT TERMS
Computer Vision, Spatiotemporal Context, Video Analysis, Action Recognition, Sparse Representation, Low-Rank Representation

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
UU	UU	UU	UU		Yun Fu
					19b. TELEPHONE NUMBER 716-645-2670

Standard Form 298 (Rev 8/98)
Prescribed by ANSI Std. Z39.18

Report Title

Final Report for Modeling Spatiotemporal Contextual Dynamics

ABSTRACT

An important problem of visual understanding is how to recognize and predict human actions or imminent events from video. The ultimate intelligent systems should be able to detect/track suspicious subjects, predict actions and events, and raise alarms for emergencies before happening. From this STIR project, we have created a new algorithmic tool set of modeling spatiotemporal contextual dynamics. For low-level and middle-level visual representation, we proposed a class of Schatten norm based discriminative metrics, locality-constrained low-rank coding, discriminative analysis by multiple principal angles, and clustering based fast low-rank approximation for large scale analysis. We also proposed decomposed contour prior and a stub feature based level set method for shape recognition in images and videos. For high-level understanding and inference, we proposed the ARMA-HMM model for early recognition of human activity and the complex temporal composition model of actionlets for activity prediction. Effectiveness and efficiency have been extensively tested for human action and activity recognition and prediction. The evaluation results and outcomes of this research have been published in 8 peer-reviewed conference proceedings along with a best paper ward, and 1 peer-reviewed journal paper.

Enter List of papers submitted or published that acknowledge ARO support from the start of the project to the date of this printing. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

Received	Paper
2012/08/08 1' 11	Ya Su, Yun Fu, Xinbo Gao, Qi Tian. Discriminant Learning Through Multiple Principal Angles for Visual Recognition, IEEE Transactions on Image Processing, (03 2012): 0. doi: 10.1109/TIP.2011.2169972

TOTAL: 1

Number of Papers published in peer-reviewed journals:

(b) Papers published in non-peer-reviewed journals (N/A for none)

Received Paper

TOTAL:

Number of Papers published in non peer-reviewed journals:

(c) Presentations

Number of Presentations: 0.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received Paper

TOTAL:

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received		Paper
2012/08/08 1	10	Zhi Yang, Yu Kong, Yun Fu. Decomposed Contour Prior for Shape Recognition, International Conference on Pattern Recognition (ICPR). 2012/11/11 00:00:00, . : ,
2012/08/08 1	9	Devansh Arpit, Gaurav Srivastava, Yun Fu. Locality-constrained Low Rank Coding for Face Recognition, International Conference on Pattern Recognition (ICPR). 2012/11/11 00:00:00, . : ,
2012/08/08 1	8	Zhenghong Gu, Ming Shao, Liangyue Li, Yun Fu. Discriminative Metric: Schatten Norm vs. Vector Norm, International Conference on Pattern Recognition (ICPR). 2012/11/11 00:00:00, . : ,
2012/08/08 1	7	Kang Li, Yun Fu. ARMA-HMM: A New Approach for Early Recognition of Human Activity, International Conference on Pattern Recognition (ICPR). 2012/11/11 00:00:00, . : ,
2012/08/08 1	6	Zhi Yang, Yu Kong, Yun Fu. Contour-HOG: A Stub Feature based Level Set Method for Learning Object Contour, British Machine Vision Conference (BMVC). 2012/09/03 00:00:00, . : ,
2012/08/08 1	5	Kang Li, Jie Hu, Yun Fu. Modeling Complex Temporal Composition ofActionlets for Activity Prediction, European Conference on Computer Vision (ECCV). 2012/10/07 00:00:00, . : ,
2012/08/08 1	4	Yuan Yao, Yuan Yao, Yun Fu. Real-Time Hand Pose Estimation From RGB-D Sensor, IEEE International Conference on Multimedia & Expo (ICME). 2012/07/09 00:00:00, . : ,
2012/08/08 1	2	Wei Chen, Ming Shao, Yun Fu. Clustering Based Fast Low-Rank Approximation for Large-Scale Graph, 2011 IEEE International Conference on Data Mining Workshops (ICDMW). 2011/12/11 00:00:00, Vancouver, BC, Canada. : ,

TOTAL: 8

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts):

(d) Manuscripts

Received		Paper
2012/08/08 1	3	Ya Su, Yun Fu, Xinbo Gao, Qi Tian. Discriminant Learning Through Multiple Principal Angles for Visual Recognition, IEEE Transactions on Image Processing (03 2012)

TOTAL: 1

Number of Manuscripts:

Books

Received	Paper

TOTAL:

Patents Submitted

Patents Awarded

Awards

The following paper was selected for the Best Paper Award in IEEE ICDM Workshop 2011.

Wei Chen, Ming Shao, and Yun Fu, “Fast Low-Rank Approximation for Large Graph,” IEEE International Conference on Data Mining (ICDM), Workshop on Large Scale Visual Analytics, pp. 787-792, 2011. (Best Paper Award)

Graduate Students

NAME	PERCENT_SUPPORTED	Discipline
Zhi Yang	0.50	
Jie Hu	0.10	
FTE Equivalent:	**0.60**	
Total Number:	**2**	

Names of Post Doctorates

NAME	PERCENT_SUPPORTED
Yu Kong	1.00
Siyu Xia	0.20
FTE Equivalent:	**1.20**
Total Number:	**2**

Names of Faculty Supported

NAME	PERCENT_SUPPORTED	National Academy Member
Yun Fu	0.10	
FTE Equivalent:	**0.10**	
Total Number:	**1**	

Names of Under Graduate students supported

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

The number of undergraduates funded by this agreement who graduated during this period: 0.00

The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:...... 0.00

The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:...... 0.00

Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):...... 0.00

Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:...... 0.00

The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense 0.00

The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields:...... 0.00

Names of Personnel receiving masters degrees

NAME

Total Number:

Names of personnel receiving PHDs

NAME

Total Number:

Names of other research staff

NAME	PERCENT_SUPPORTED

FTE Equivalent:
Total Number:

Sub Contractors (DD882)

Inventions (DD882)

Scientific Progress

See Attachment

Technology Transfer

DEPARTMENT OF THE ARMY
UNITED STATES ARMY RESEARCH LABORATORY
ARMY RESEARCH OFFICE
P. O. BOX 12211
RESEARCH TRIANGLE PARK NC 27709-2211

REPLY TO
ATTENTION OF

Final Report: Scientific Progress and Accomplishments

Title: Modeling Spatiotemporal Contextual Dynamics with Sparse-Coded Transfer Learning

Proposal No: 59118-CS-II

Grant No: U. S. Army Research Office STIR under W911NF-11-1-0365

PI: **Yun Fu**, Assistant Professor, SUNY at Buffalo

1. Statement of the Problem Studied

An important problem of visual understanding is how to recognize and predict human actions or imminent events from video. Figure 1 shows some real-world army scenarios in the battlefield environment, where intelligent unmanned ground vehicles or robots can provide real-time surveillance and monitoring. For example, if a kid carrying a box wants to give it to a soldier, shall we predict it is a bomb? If someone coming with a bag but returns without it, shall we predict it is an abandon of the bag? If a guy runs to approach a soldier, shall we predict the soldier is in dangerous? If some mobs pick up something, shall we predict they will throw to the military equipment? The PI envisions the ultimate intelligent systems can detect/track suspicious subjects, predict actions and events, and raise alarms for emergencies before happening. ***The research objective of this proposal is to create new theoretical methodologies of modeling spatiotemporal contextual dynamics***. As a short term innovative research project, it particularly focused on human action recognition based on the proposed new models.

(a) Carry and give? (b) Abandon and leave? (c) Run and approach? (c) Pick up and throw?

Figure 1: Examples of real-world army scenarios of action recognition and prediction through video surveillance.

With the state-of-the-art techniques, human detection and tracking can be achieved reliably in some systems under well-constrained sensing conditions using boosted low-level visual features. However, in the mid-level, human action and event detection and recognition are still open problems due to the difficulty of tracking rigid parts of articulated objects (such as human arms, legs, head and torch) and inferring the accurate dynamics. Specifically, most existing systems can only deal with constrained actions of "Noun" for a single subject. It is difficult to learn a robust generative model that can capture all possible shapes of human body with large variances and occlusions. Appearance based discriminative method models the whole appearance variances without measuring the large freedom of joint angles. However, such approaches only work in the "clean" data, with uniform background, non-occlusion, subject collaborative performance, perfect spatiotemporal segmentation, and well-defined action context. When applied to real-world data without any constraints, such as Figure 1, there is still no reliable method that can perform high-performance human action or event recognition in complex environments, and even further to support the prediction of imminent actions or events.

2. Summary of Results

From this STIR project, we have created a new algorithmic tool set of modeling spatiotemporal contextual dynamics [1-9]. For low-level and middle-level visual representation, we proposed a class of

Schatten norm based discriminative metrics, locality-constrained low-rank coding, discriminative analysis by multiple principal angles, and clustering based fast low-rank approximation for large scale analysis. We also proposed decomposed contour prior and a stub feature based level set method for shape recognition in images and videos. For high-level understanding and inference, we proposed the ARMA-HMM model for early recognition of human activity and the complex temporal composition model of actionlets for activity prediction. Effectiveness and efficiency have been extensively tested for human action and activity recognition and prediction. The evaluation results and outcomes of this research have been published in 8 peer-reviewed conference proceedings along with a best paper ward, and 1 peer-reviewed journal paper. For more details of the research outcomes and accomplishments, please refer to the attached publications [1-9]. This report will highlight the details in [3,8].

2.1 Modeling Complex Temporal Composition

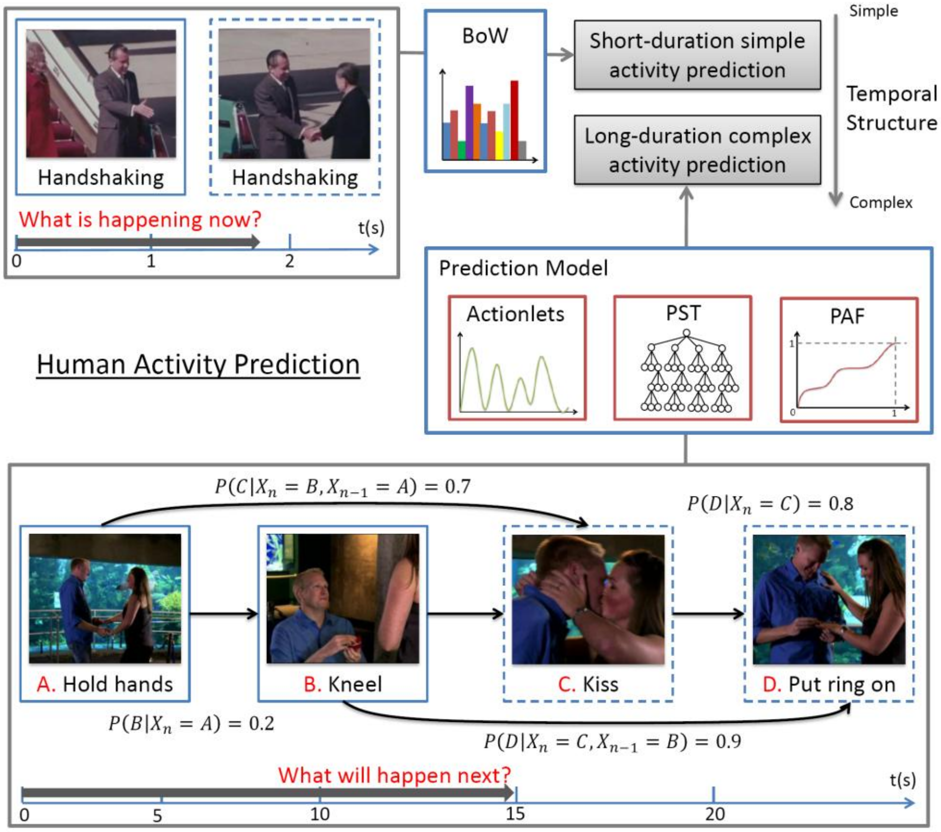

Figure 2: Frameworks for two categories of activity prediction problems: (1) short-duration simple activity prediction (e.g. "handshaking"), and (2) long-duration complex activity prediction (e.g. "propose marriage"). The first problem can be solved in the classic bag-of-words paradigm. Our approach aims to solve the second problem.

In particular, we propose a novel framework, shown in Figure 2, for predicting long-duration complex activity by discovering the causal relationships between constituent actions and predictable characteristic of the activities. The key of our approach is to utilize the observed action units as context to predict the next possible action unit, or predict the intension and effect of the whole activity. It is thus possible to make prediction with meaningful earliness and have the machine vision system provide a time-critical reaction. We represent complex activity as sequences of discrete action units, which have specific semantic meanings and clear time boundaries. To ensure a good discretization, we propose a novel temporal segmentation method for action units by discovering the regularity of motion velocity.

And the key contribution of this work is the idea that causality of action units can be encoded as a Probabilistic Suffix Tree (PST) with variable temporal scale, while the predictability can be characterized by a Predictive Accumulative Function (PAF) learned from information entropy changes along every stage of activity progress. In order to test the efficacy of our method, we introduce a new dataset that focuses on complex activity in tennis game. Our method aims to answer the challenging question: “can we predict who will win?". Also we test our method on another benchmark dataset about daily indoor living activity. Our algorithm shows very promising results. The generalization capability of this new model is inductive enough to extend the application to ARMY video data.

2.2 Dataset

Our prediction model can be applied to a variety of human activities. The key requirement is that the activity should have multiple steps where each step constitutes a meaningful action unit. Without loss of generality, we choose two datasets with significant different temporal structure complexity. First, we collect real world video for tennis games between two top male players from YouTube, see Figure 3. Each point with an exchange of several strokes is considered as an activity instance, which involves two agents. In total, we intercepted 160 video clips for 160 points from a 4 hour game. Then we separate them into two categories of activity, where 80 clips are winning points and 80 clips are losing points with respect to a speci_c player. So our prediction problem on this dataset becomes an interesting question: “can we predict who will win?". Since each point consists of sequence of action units with length ranging from one to more than twenty, tennis game has a high-level temporal structure complexity in terms of both variance and order. Second, we choose Maryland Human-Object Interactions (MHOI) dataset [3], which consists of 5 annotated activities: answering a phone call, making phone call, drinking water, lighting a flash, pouring water into container. These activities have about 3 to 5 action units each. And constituent action units share similar human movements: 1) reaching for an object of interest, 2) grasping the object, 3) manipulating the object, and 4) put back the object. For each activity, we have 9 or 10 video samples. And there are 44 video clips in total.

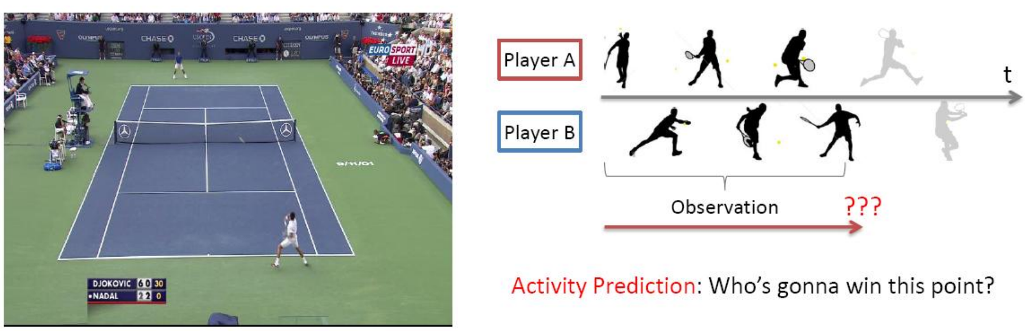

Figure 3: YouTube tennis game dataset.

2.3 Temporal Decomposition and Video Representation

Temporal decomposition is the first key step for our representation of complex activity. It is to find the frame indices that can segment a long sequence of human activity video into multiple meaningful atomic actions. We call these atomic actions actionlets. We found that the velocity changes of human actions have similar periodic regularity. The specific method has following several steps: 1) Use Harris Corner to find significant key points; 2) Use Lucas-Kanade (LK) optical flow to generate the trajectories for key points; 3) For each frame, accumulate the trajectories/tracks at these points to get a velocity magnitude.

Based on accurate temporal decomposition results, we can easily cluster actionlet into meaningful groups so that each activity can be represented by a sequence of actionlets in a syntactic way. A variety

of video descriptors can be used here as long as it can provide a discriminative representation for the actionlets.

2.4 Activity Prediction Model

Causality is an important cue for human activity prediction. So automatic acquisition of causality from sequential actionlets becomes the key. Variable order Markov Model (VMM) is a category of algorithms for prediction of discrete sequences. It suits the activity prediction problem well, because it can capture both large and small order Markov dependencies based on training data. Therefore, it can encode richer and more flexible causal relationships. Here, we model complex human activity as a Probabilistic Suffix Tree (PST) which implements the single best L-bounded VMM (VMMs of degree L or less) in a fast and efficient way. To characterize the predictability of activities, we formulate a Predictive Accumulative Function (PAF). We want to depict the predictable characteristic of a particular activity. For example, “tennis game" is a late-predictable problem in the sense that when we observed a long sequence of actionlets performed by two players, the last several strokes will strongly impact the winning or losing results. In contrast, “drinking water" is an early predictable problem, since as long as we observed the first actionlet “grabbing a cup", we probably can guess the intention. So different activities always have quite different PAFs. In our model, PAF can be learned automatically from the training data. And later when do prediction, we use PAF to weight the observed patterns in every stage of ongoing sequence.

2.5 Evaluation Results

We test the ability of our approach to predict human activities with middle-level temporal complexity on MHOI dataset. Samples in MHOI dataset are about daily activities (e.g. “making phone call"). This type of activity usually consists of 3 to 5 actionlets and lasts about 5 to 8 seconds, so we call it middle-level complex activity. In this dataset, each category has 9 or 10 samples. For a particular activity, we use all the samples in that category as positive set, and randomly select equal number of samples from remaining categories as negative set. Then we fit the prediction task into the context of supervised classification problem. To train a prediction model, we construct an order 5-bounded PST and fit a PAF respectively. To evaluate the prediction accuracy, we use “leave-one-out" method. Since the sample number is relatively small, we repeat our experiments 10 times for each activity and average the performance. In addition, we implemented several previous human activity prediction approaches to compare them with our method. Three types of previous prediction model using the same features were implemented: (1) Dynamic Bag-of-Words model, (2) Integral Bag-of-Words model, and (3) a basic SVM-based approach.

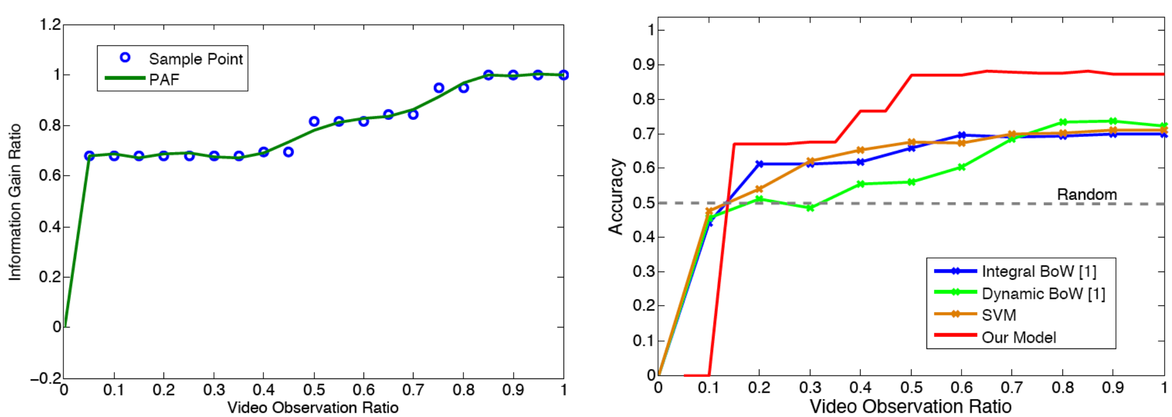

Figure 4: Activity prediction results on MHOI dataset.

Figure 4 illustrates the process of fitting PAF from training data. It shows that daily activities such as examples from MHOI dataset are early predictable. That means the semantic information at early stage strongly exposes the intension of the whole activity. The results of the implemented 4 methods are averaged over 5 activities. The proposed method has great advantages over other methods. For example, after half video observed (about 2 actionlets), our model is able to make a prediction with the accuracy of 0.9. We then test our model at high-level temporal complexity activities on the tennis game dataset.

We aim to test the ability of our model to leverage the temporal structure of human activity. Each sample video in the tennis game dataset is corresponding to a point which consists of a sequence of actionlets (strokes). The length of actionlet sequence of each point can vary from 1 to more than 20. So the duration of some sample videos may as long as 30 seconds. We group samples into two categories, winning and losing, with respect to a specific player. Overall, we have 80 positive and 80 negative samples respectively. Then a 6-bounded PST and a PAF are trained from data to construct the prediction model. And the same "leave-one-out" method is used for evaluation. Table 1shows detailed comparison of 4 methods on two datasets. Random guess is 0.5, therefore the other 3 methods actually perform random guess on tennis game.

Table 1. Performance comparisons on two datasets.

Methods	Tennis Game Dataset					MHOI dataset				
	20% observed	40% observed	60% observed	80% observed	100% observed	20% observed	40% observed	60% observed	80% observed	100% observed
Integral BoW [1]	0.47	0.44	0.53	0.47	0.51	0.61	0.62	0.70	0.69	0.70
Dynamic BoW [1]	0.53	0.55	0.49	0.44	0.48	0.51	0.56	0.60	0.73	0.72
SVM	0.56	0.52	0.51	0.48	0.49	0.54	0.65	0.67	0.70	0.71
Our Model	**0.59**	**0.64**	**0.65**	**0.65**	**0.70**	**0.67**	**0.77**	**0.87**	**0.88**	**0.87**

2.6 Conclusion

Our approach is a general framework for activity prediction. It can be integrated with any sequential decomposition methods of complexity activity with flexible actionlets granularity. It is a brand new method customized to the prediction problem. Since activity classification and activity prediction are quite different problems, it is inappropriate to adopt similar bag-of-words paradigm. All the experimental results validate the advantages of utilizing causality and predictability as prediction driving force, which inspires us to follow this philosophy principal when design new activity prediction techniques. Compared with the rare existing work for activity prediction, our approach outperforms their method by a large margin on both accuracy and earliness. To our best knowledge, the proposed model is the only one that can predict on high-level complex activities.

3. Scientific Significance

Conventional approaches, usually focusing on dealing with a sub-problem of spatiotemporal composition, fail to model such dynamics and structural nature of motions for the purpose of action recognition and prediction in unconstrained scenarios. As a rich source of dynamic context, such unconstrained army data can be modeled through spatiotemporal contextual dynamics in a large scale. By studying novel methodologies of visual pattern extraction in a mathematically coherent learning framework, the conducted research is the first complete model for action prediction in unconstrained visual media. Such progresses will significantly advance the visual intelligence field and contribute to the accomplishment of the Army's mission.

4. Future Research Plans

The theoretical contribution of this research will pave the foundation for novel techniques in solving important problems of visual understanding and large-scale visual analytics. Such research endeavor is sustainable and can go well beyond the 9-month scope, which is the PI's long-term career goal. Leveraged by this grant, the PI will submit an ARO Young Investigator Program proposal with a title of “Video Understanding under Uncertainty by Low-Rank Analytics”. The leveraged ARO YIP proposal is a concrete example of future research plan within the PI’s key research interests.

5. Bibliography

[1] Wei Chen, Ming Shao, and Yun Fu, “Fast Low-Rank Approximation for Large Graph,” IEEE International Conference on Data Mining (ICDM), Workshop on Large Scale Visual Analytics, pp. 787-792, 2011. (**Best Paper Award**)

[2] Ya Su, Yun Fu, Xinbo Gao and Qi Tian, “Discriminant Learning through Multiple Principal Angles for Visual Recognition,” IEEE Transactions on Image Processing (T-IP), Volume: 21 , Issue: 3, Page(s): 1381 - 1390, 2012.

[3] Kang Li, Jie Hu, Yun Fu. Modeling Complex Temporal Composition ofActionlets for Activity Prediction, European Conference on Computer Vision (ECCV). 07-OCT-12.

[4] Zhi Yang, Yu Kong, Yun Fu. Contour-HOG: A Stub Feature based Level Set Method for Learning Object Contour, British Machine Vision Conference (BMVC). 03-SEP-12.

[5] Zhi Yang, Yu Kong, Yun Fu. Decomposed Contour Prior for Shape Recognition, International Conference on Pattern Recognition (ICPR). 11-NOV-12.

[6] Devansh Arpit, Gaurav Srivastava, Yun Fu. Locality-constrained Low Rank Coding for Face Recognition, International Conference on Pattern Recognition (ICPR). 11-NOV-12.

[7] Zhenghong Gu, Ming Shao, Liangyue Li, Yun Fu. Discriminative Metric: Schatten Norm vs. Vector Norm, International Conference on Pattern Recognition (ICPR). 11-NOV-12.

[8] Kang Li, Yun Fu. ARMA-HMM: A New Approach for Early Recognition of Human Activity, International Conference on Pattern Recognition (ICPR). 11-NOV-12.

[9] Yuan Yao, Yuan Yao, Yun Fu. Real-Time Hand Pose Estimation From RGB-D Sensor, IEEE International Conference on Multimedia & Expo (ICME). 09-JUL-12.